Semi-supervised Quality Evaluation of Colonoscopy Procedures

I. Kligvasser, G. Leifman, R. Goldenberg, E. Rivlin, and M. Elad

Verily Life Sciences

Abstract. Colonoscopy is the standard of care technique for detecting and removing polyps for the prevention of colorectal cancer. Nevertheless, gastroenterologists (GI) routinely miss approximately 25% of polyps during colonoscopies. These misses are highly operator dependent, influenced by the physician skills, experience, vigilance, and fatigue. Standard quality metrics, such as Withdrawal Time or Cecal Intubation Rate, have been shown to be well correlated with Adenoma Detection Rate (ADR). However, those metrics are limited in their ability to assess the quality of a specific procedure, and they do not address quality aspects related to the style or technique of the examination. In this work we design novel online and offline quality metrics, based on visual appearance quality criteria learned by an ML model in an unsupervised way. Furthermore, we evaluate the likelihood of detecting an existing polyp as a function of quality and use it to demonstrate high correlation of the proposed metric to polyp detection sensitivity. The proposed online quality metric can be used to provide real time quality feedback to the performing GI. By integrating the local metric over the withdrawal phase, we build a global, offline quality metric, which is shown to be highly correlated to the standard Polyp Per Colonoscopy (PPC) quality metric.

1 Introduction

Screening for colorectal cancer is highly effective, as early detection is within reach, making this disease one of the most preventable. Today's standard of care screening method is optical colonoscopy, which searches the colon for mucosal abnormalities, such as polyps. However, performing a thorough examination of the entire colon surface using optical colonoscopy is challenging, which may lead to a lower polyp detection rate. Recent studies have shown that approximately 25% of polyps are routinely missed during colonoscopies [1].

The success (diagnostic accuracy) of a colonoscopy procedure is highly operator dependent. It varies based on the performing physician skills, experience, vigilance, fatigue, and more. To ensure high procedure quality, various quality metrics are measured and monitored. E.g., the Withdrawal Time (time from the colonoscope reaching cecum to removal of the instrument from the patient) metric was shown to be highly correlated to Adenoma Detection Rate (ADR) [6, 13, 15, 16, 17, 18]. Another quality metric – Cecal Intubation Rate (proportion of colonoscopies in which the cecum is intubated) – is considered important to ensure good colon coverage.

Most of these existing metrics are relatively easy to compute, but can provide only limited data on the quality of a specific procedure, and are typically used aggregatively for multiple sessions. Some studies [14] suggest that there are other factors that impact the polyp detection rate. For example, one may wish to distinguish between a good and bad colonoscope motion patterns, or assess the style of the examination. The hypothesis is that a better inspection style yields more informative visual input, which results in a better diagnostic accuracy.

In this work we propose a novel quantitative quality metric for colonoscopy, based on the automatic analysis of the induced video feed. This metric is computed locally in time, measuring how informative and helpful for colon inspection a local video segment is. As this instantaneous quality is very subjective and difficult to formulate, human annotation is problematic and ill-defined. Instead, we let an ML model build a meaningful visual data representation in a fully unsupervised way, and use it to construct a metric highly correlated with the clinical outcome. First, we learn visual representations of colonoscopy video frames using contrastive self-supervised learning. Then, we perform cluster analysis on these representations and construct a learned aggregation of these cluster assignments, bearing a strong correlation with polyp detection, which can serve as an indicator for "good-quality" video segments.

While the proposed approach resembles the one proposed in [7], the addressed problems are markedly different, as [7] does phase detection in colonoscopy. There are other works aiming to learn frame representations in colonoscopy videos, However, those descriptors are usually associated with polyps, and used for polyp related tasks - tracking, re-identification [3, 19], optical biopsy [20], etc.

By measuring the duration of good quality video segments over the withdrawal phase of the procedure, we derive a new offline colonoscopy quality metric. We show that this measure is strongly correlated to the Polyps Per Colonoscopy (PPC) quality metric. Moreover, we show how the real-time measurement of the quality of a colonoscopy procedure can be used to evaluate the likelihood of detecting a polyp at any specific point in time during the procedure.

2   Method

Our goal is to learn a colonoscopy quality metric through the identification of temporal intervals in which effective polyp detection is possible. We start by learning the colonoscopy video frame embedding using self-supervised learning, followed by a cluster analysis. Using those clusters, we learn a "good" frame classifier, which then serves as the basis for both global (offline) and local (online) quality metrics. The end-to-end framework is described in the following sections, and illustrated in Fig. 1.

2.1   Frame Encoding

We start from learning visual representations of colonoscopy frames using contrastive learning. We use SimCLR [4], which maximizes the agreement between

Fig. 1: **Method overview.** *(Left)* Two augmented views for each frame are used to train the encoder and the projection head using contrastive learning. *(Right top)* Feature representations are directly clustered into semantically meaningful groups using K-means. *(Right middle)* Learning clusters' associations. *(Right bottom)* At inference time, cluster attributes are leveraged for quality metric evaluation.

representations of two randomly augmented versions of the same frame, while pushing away the representations of other frames (see Fig. 1). Specifically, frame x_i is randomly augmented, resulting in two correlated views, x_i^1 and x_i^2, considered as a positive pair. These views are fed to an encoder $f_\theta(\cdot)$ and projection layer $g_\phi(\cdot)$, yielding the embedding vector $z_i^a = g_\phi(f_\theta(x_i^a))$ $(a = 1, 2)$. Given a batch of N frames, the contrastive loss referring to the i-th frame is given by

$$\ell(z_i^1, z_i^2) = -log \frac{exp(sim(z_i^1, z_i^2)/\tau)}{\sum_{k \neq i} \sum_{a=1}^2 \sum_{b=1}^2 exp(sim(z_i^a, z_k^b)/\tau)}, \quad (1)$$

where τ is a temperature parameter, and sim is the cosine similarity defined as $sim(u, v) = u^T v / \|u\| \|v\|$. We use ResNet-RS50 [2] for the encoder and a simple MLP with one hidden layer for the projection layer, as suggested in [4].

Our training data consists of $1M$ frames randomly sampled from 2500 colonoscopy videos. Since the designed metric is supposed to be used for predicting the chance of detecting a polyp, it is not expected to be used on frames where the polyp is detected or treated. Therefore, we exclude such frames from the training set, by detecting them automatically using off-the-shelf polyp and surgical tool detectors [9, 10, 11].

For augmentation we use standard geometric transformations (resize, rotation, translation), color jitter, and the Cutout [5] with the Gaussian noise filling.

Fig. 2: **T-SNE plot of frame embeddings.** K-means clusters are color coded.

2.2 Frame Clustering

The second step in our scheme is clustering the learned representations[1] $f_\theta(x_i)$ into $K(=10)$ clusters using k-means [12]. While the standard k-means does a hard assignment of each frame to its corresponding cluster, we use a soft alternative based on the distance between the frame descriptor to cluster centers. Namely, we define the probability of the i-th frame to belong to the k-th cluster by

$$r_{i,k} = Prob(f_\theta(x_i) \in k) \sim \left[\frac{1}{\|f_\theta(x_i) - c_k\|_2^2} \right]^\alpha \quad \text{for} \quad k = 1, 2, \ \dots \ , K, \quad (2)$$

where $\{c_k\}_{k=1}^K$ are the cluster centers, $\alpha = 16$, and $\{r_{i,k}\}_{k=1}^K$ are normalized to sum to 1. Figure 2 shows the t-SNE projection of frame embeddings with k-means clusters color coded. Interestingly, the samples are clustered into relatively compact, meaningful groups. Figure 3 presents a random selection of frames from each cluster. One can see that clusters $1, 2$ and 7 contains inside-body informative frames. In contrast, clusters $0, 3, 4, 5, 6, 8$ and 9 contain non-informative outside-body and inside-body frames. Please see the SM for more visual examples.

2.3 Online (Local) Quality Metric

Based on the learned frame embeddings and clusters, we now design an online (local) quality metric. As our objective is to link the visual appearance to polyp detection, we will learn a metric that tries to predict one from the other. Namely,

---

[1] Note that the projection head $g_\phi(\cdot)$ is omitted from here on.

Fig. 3: **Clusters visualization.** Random selection of frames from each cluster.

we learn a function $Q(\cdot)$ that maps frame x_i appearance encoded by the vector $\{r_{i,k}\}_{k=1}^K$ (see Eq. 2) to the chance of detecting a polyp in the following frames.

More precisely, we average the $\{r_{i,k}\}_{k=1}^K$ over a video segment of 10 sec to get $\{\overline{r_{i,k}}\}_{k=1}^K$, and train a binary classifier $Q(\{\overline{r_{i,k}}\}_{k=1}^K)$ to predict the detection of a polyp in the following 2 sec.

The training set for the classifier is built from a set of 2243 colonoscopy videos annotated for the location of polyps. 1086 intervals of 10 seconds before the appearance of polyps are sampled from the training set as positive samples, and another 1086 random intervals sampled as negative samples. The $Q(\cdot)$ is implemented as a binary classifier with a single linear layer and trained with Adam optimizer [8] for 500 epochs, using a batch size of 64.

While the $Q(\cdot)$ achieves only a mediocre classification (i.e. polyp detection prediction) accuracy of 64% on the test-set (indeed, it is very difficult to predict a detection of a polyp when it is not known that the polyp is there), we will show in the following sections that it can still be used as a quality metric.

2.4 From Quality Metric to the Chance to Detect a Polyp

We would like to assess the chance of detecting a polyp (if it exists) at a certain time point t as a function of the procedure quality Q in the preceding time interval $[t - \Delta t, t]$. Let us denote the event of having a polyp in the colon at time t as E ("exists"), and the event of detecting it as D ("detected"). For this analysis we will treat the quality metric Q from the previous section, as a random variable in the range $[0, 1]$ measuring the quality of the procedure in the time interval $[t - \Delta t, t]$.

We are interested to estimate the following probability:

$$P(D|E,Q) = \frac{P(E,Q|D)P(D)}{P(E,Q)} = \frac{P(Q|D)P(E|Q,D)P(D)}{P(E,Q)}, \tag{3}$$

representing the chance of detecting a polyp if it exists as a function of quality. In the above, the first equality uses the Bayes rule, and the second exploits the chain probability relationship. We know that physicians rarely mistake a non-polyp for a polyp, implying that $P(E|Q, D) \approx 1$. Then, assuming the independence between the existence of the polyp (E) and the quality of the procedure (Q), Eq. 3 becomes

$$P(D|E, Q) \approx \frac{P(Q|D)P(D)}{P(Q)P(E)} \qquad (4)$$

As mentioned above, the incidence of polyp detection false alarms in colonoscopy is negligible, hence the ratio $P(D)/P(E)$ can be interpreted as the average polyp detection rate/sensitivity (PDS). From the literature, we know that polyp miss-rate in colonoscopy is about $20 - 25\%$ [1]. Hence, $P(D)/P(E)$ can be approximated as $0.75 - 0.8$, regardless of Q.

Therefore, to compute $P(D|E, Q)$, all we need to do is approximate $P(Q)$ and $P(Q|D)$. This can be done empirically by estimating the distribution of Q in random intervals and in intervals preceding polyps for $P(Q|D)$.

2.5   Offline Quality Metric (Post-Procedure)

We would like to design an offline quality indicator based on the above online measure Q. We define the following quality metric by integrating Q over the entire withdrawal phase,

$$Q_{\text{Offline}} = \sum_{i \in \text{withdrawal}} Q\left(\{r_{i,k}\}_{k=1}^{K}\right). \qquad (5)$$

3   Experiments

3.1   Online Quality Metric Evaluation

We would like to evaluate how relevant the proposed online quality metric Q is to the ability of detecting polyps. We do that by estimating the likelihood of detecting an existing polyp $P(D|E, Q)$ as a function of Q. The higher the correlation between Q and $P(D|E, Q)$, the better Q is as a local colonoscopy quality metric.

As discussed above $P(D|E, Q) \propto P(Q|D)/P(Q)$. Both $P(Q|D)$ and $P(Q)$ can be estimated empirically: For $P(Q)$ we build a 10-bin histogram of Q measured in 543 randomly chosen colonoscopy video segments $10sec$ long. The same is done for $P(Q|D)$, but with 543 video segments preceding a polyp.

The estimated $P(D|E, Q)$ is depicted in Figure 4. As one can see, the proposed quality metric Q correlates very well with the polyp detection sensitivity (PDS). Q can be computed online and provided as a real time feedback to the physician during the procedure.

Fig. 4: The likelihood of detecting an existing polyp in a short video segment as a function of local quality metric Q.

3.2 Offline Quality Metric Evaluation

We would like to evaluate the effectiveness of the proposed offline quality metric Q_{Offline} in predicting the polyp detection sensitivity.

To do so, we compute Q_{Offline} for 500 annotated test set colonoscopies. We sort the cases in the increasing order of Q_{Offline}, and split them into 5 bins - 100 cases each, from lower Q_{Offline} to higher. For each bin we compute the average Polyps Per Colonoscopy (PPC) metric. The resulting historgram is shown in Fig. 5(Left). One can observe a strong correlation between the Q_{Offline} and the PPC metric.

Fig. 5(Right) shows the distribution of procedures with (red) and without detected polyps (blue), as the function of Q_{Offline}. One can see that higher Q_{Offline} are more likely to correspond to procedures with detected polyps.

The evaluations above suggest that the proposed quality metric Q_{Offline} is highly correlated to polyp detection sensitivity (PPS). It is important to note that high Q_{Offline} for any specific procedure does not mean that there is a high chance of finding a polyp in that procedure, as we don't know if there are any polyps there and how many. What it does mean, is that if there is a polyp, there is a high chance it will be detected.

4 Conclusion

We proposed novel online and offline colonoscopy quality metrics, computed based on the visual appearance of frames in colonoscopy video. The quality criteria for the visual appearance were automatically learned by an ML model in an unsupervised way.

Using a Bayesian approach, we developed a technique for estimating the likelihood of detecting an existing polyp as a function of the proposed local quality

Fig. 5: **Q_{Offline} during the withdrawal phase.** (Left) The relationship between the proposed offline quality measure and the actual number of polyps detected, when Q_{Offline} observations are divided into five equal-sized groups. (Right) Procedures with high Q_{Offline} values are likely to have polyps.

metric. We used this likelihood estimation to demonstrate the correlation between the local quality metric and the polyp detection sensitivity. The proposed local metric can be computed online to provide a real time quality feedback to the performing physician.

Integrating the local metric over the withdrawal phase yields a global, offline quality metric. We show that the offline metric is highly correlated to the standard Polyps Per Colonoscopy (PPC) quality metric.

As the next step, we would like to estimate the impact of the proposed real time quality feedback on the quality of the procedure, e.g. by measuring its impact on the Adenoma Detection Rate (ADR) in a prospective study.

Bibliography

[1] Ahn, S.B., Han, D.S., Bae, J.H., Byun, T.J., Kim, J.P., Eun, C.S.: The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. Gut and liver **6**(1), 64 (2012)

[2] Bello, I., Fedus, W., Du, X., Cubuk, E.D., Srinivas, A., Lin, T.Y., Shlens, J., Zoph, B.: Revisiting resnets: Improved training and scaling strategies. Advances in Neural Information Processing Systems **34**, 22614–22627 (2021)

[3] Biffi, C., Salvagnini, P., Dinh, N.N., Hassan, C., Sharma, P., Cherubini, A.: A novel ai device for real-time optical characterization of colorectal polyps. NPJ digital medicine **5**(1), 84 (2022)

[4] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)

[5] DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)

[6] Fatima, H., Rex, D.K., Rothstein, R., Rahmani, E., Nehme, O., Dewitt, J., Helper, D., Toor, A., Bensen, S.: Cecal insertion and withdrawal times with wide-angle versus standard colonoscopes: a randomized controlled trial. Clinical Gastroenterology and Hepatology **6**(1), 109–114 (2008)

[7] Kelner, O., Weinstein, O., Rivlin, E., Goldenberg, R.: Motion-based weak supervision for video parsing with application to colonoscopy. arXiv preprint arXiv:2210.10594 (2022)

[8] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[9] Lachter, J., Schlachter, S.C., Plowman, R.S., Goldenberg, R., Raz, Y., Rabani, N., Aizenberg, N., Suissa, A., Rivlin, E.: Novel artificial intelligence–enabled deep learning system to enhance adenoma detection: a prospective randomized controlled study. iGIE (2023)

[10] Leifman, G., Aides, A., Golany, T., Freedman, D., Rivlin, E.: Pixel-accurate segmentation of surgical tools based on bounding box annotations. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 5096–5103. IEEE (2022)

[11] Livovsky, D.M., Veikherman, D., Golany, T., Aides, A., Dashinsky, V., Rabani, N., Shimol, D.B., Blau, Y., Katzir, L., Shimshoni, I., et al.: Detection of elusive polyps using a large-scale artificial intelligence system (with videos). Gastrointestinal Endoscopy **94**(6), 1099–1109 (2021)

[12] Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982)

[13] Sanchez, W., Harewood, G.C., Petersen, B.T.: Evaluation of polyp detection in relation to procedure time of screening or surveillance colonoscopy. Official journal of the American College of Gastroenterology| ACG **99**(10), 1941–1945 (2004)

[14] Sawhney, M.S., Cury, M.S., Neeman, N., Ngo, L.H., Lewis, J.M., Chuttani, R., Pleskow, D.K., Aronson, M.D.: Effect of institution-wide policy of colonoscopy withdrawal time more than 7 minutes on polyp detection. Gastroenterology **135**(6), 1892–1898 (2008)

[15] Shaukat, A., Rector, T.S., Church, T.R., Lederle, F.A., Kim, A.S., Rank, J.M., Allen, J.I.: Longer withdrawal time is associated with a reduced incidence of interval cancer after screening colonoscopy. Gastroenterology **149**(4), 952–957 (2015)

[16] Shine, R., Bui, A., Burgess, A.: Quality indicators in colonoscopy: an evolving paradigm. ANZ journal of surgery **90**(3), 215–221 (2020)

[17] Simmons, D.T., Harewood, G.C., Baron, T.H., Petersen, B.T., Wang, K.K., Boyd-Enders, F., Ott, B.J.: Impact of endoscopist withdrawal speed on polyp yield: implications for optimal colonoscopy withdrawal time. Alimentary pharmacology & therapeutics **24**(6), 965–971 (2006)

[18] Vavricka, S.R., Sulz, M.C., Degen, L., Rechner, R., Manz, M., Biedermann, L., Beglinger, C., Peter, S., Safroneeva, E., Rogler, G., et al.: Monitoring colonoscopy withdrawal time significantly improves the adenoma detection rate and the performance of endoscopists. Endoscopy **48**(03), 256–262 (2016)

[19] Yu, T., Lin, N., Zhang, X., Pan, Y., Hu, H., Zheng, W., Liu, J., Hu, W., Duan, H., Si, J.: An end-to-end tracking method for polyp detectors in colonoscopy videos. Artificial Intelligence in Medicine **131**, 102363 (2022)

[20] van der Zander, Q.E., Schreuder, R.M., Fonollà, R., Scheeve, T., van der Sommen, F., Winkens, B., Aepli, P., Hayee, B., Pischel, A.B., Stefanovic, M., et al.: Optical diagnosis of colorectal polyp images using a newly developed computer-aided diagnosis system (cadx) compared with intuitive optical diagnosis. Endoscopy **53**(12), 1219–1226 (2021)